

The bias–variance decomposition

Ten problems on bias and variance: the mean squared error of an estimator split into squared bias plus variance, the sample mean and a shrunken mean that beats it, the three-term decomposition of expected test error with the noise floor, the one-dimensional least-squares estimator and its ridge version with bias, variance and the λ that minimises their sum, how the two terms move with λ , why averaging models cuts variance but not bias and where correlation puts the floor, the variance estimator with the smallest error, and nearest-neighbour regression where k trades bias against variance, with worked solutions and the mistakes that prefer unbiased by reflex, apply the decomposition to training error, guess the ridge variance by analogy, forget the noise term, or divide the variance of correlated models by their number.

BEFORE YOU START

An estimator can be wrong in two ways: on average, which is bias, and from one dataset to the next, which is variance. Mean squared error is the sum of the square of the first and the second, and almost every design decision in machine learning, regularisation strength, model size, how many neighbours, how many models to average, moves one term up and the other down. These ten problems prove the decomposition for a scalar estimator and for a prediction at a point, where a third term, the noise, appears; work it out for the sample mean and show that shrinking it towards zero lowers the error; do the same for the one-dimensional least-squares slope and its ridge version, finding the regularisation that minimises the total; show how the two terms move with the regulariser; compute what averaging models does to each term; find the variance estimator with the smallest error, which is neither the unbiased one nor the maximum-likelihood one; and end with nearest-neighbour regression, where the number of neighbours is the knob. The five mistakes at the end are the ones that sound like good statistics: unbiased taken to mean best, the decomposition applied to the training set, a ridge variance written by analogy, the noise term dropped, and the variance of an average divided by the number of models when the models are correlated.

- An estimator $\hat{\theta}$ is a function of a random sample, so it is a random variable; θ is the fixed quantity it estimates. Its bias is $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$, its variance is $\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$ and its mean squared error is $\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$ (the maximum-likelihood page). $\mathbb{E}$ and Var are over the sample; the variance page's rules $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, $\text{Var}(aX) = a^2 \text{Var}(X)$ and $\text{Var}(\sum_n X_n) = \sum_n \text{Var}(X_n)$ for independent X_n are used throughout.

- A sample $x_1, \dots, x_N$ is independent and identically distributed (i.i.d.) with mean μ and variance σ^2 ; $\bar{x} = \frac{1}{N} \sum_n x_n$ is the sample mean and $S = \sum_n (x_n - \bar{x})^2$. For Gaussian data, S/σ^2 has a chi-square distribution with $N - 1$ degrees of freedom, whose mean is $N - 1$ and variance $2(N - 1)$; this is taken as given.
- Regression data are $y = f(x) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \sigma^2$, and the noise on different points independent. A learner trained on a dataset D produces the prediction $\hat{f}_D(x)$; $\mathbb{E}_D$ is the expectation over datasets and Var_D the variance across them. A fresh test point has its own noise ε , independent of D .
- The one-dimensional linear model has fixed inputs $x_1, \dots, x_N$, true slope w , observations $y_n = wx_n + \varepsilon_n$, and $s = \sum_n x_n^2$. Least squares gives $\hat{w} = \sum_n x_n y_n / s$ and ridge with penalty $\lambda \geq 0$ gives $\hat{w}_\lambda = \sum_n x_n y_n / (s + \lambda)$, the minimiser of $\sum_n (y_n - wx_n)^2 + \lambda w^2$ (the regression page).
- $\mathbf{1}$ is the all-ones vector, Σ a covariance matrix and ρ a correlation coefficient. k -nearest-neighbour regression predicts $\hat{f}(x) = \frac{1}{k} \sum_{j \in N_k(x)} y_j$, the mean of the k training responses nearest to x .

Builds on: Variance, covariance and correlation, Maximum likelihood estimation: derivations by hand

Problems

1. Show that for any estimator $\hat{\theta}$ of a fixed θ , $\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$. difficulty ·

2. For an i.i.d. sample with mean μ and variance σ^2 , compute the bias, difficulty · ·
variance and MSE of the sample mean $\bar{x}$. Then do the same for the shrunk estimator $c\bar{x}$ with $0 \leq c \leq 1$, find the c^* that minimises its MSE, and evaluate for $\mu = 1$, $\sigma^2 = 1$, $N = 4$.

3. For a fixed test input x with $y = f(x) + \varepsilon$ and a learner $\hat{f}_D$, show that difficulty · ·
 $\mathbb{E}_{D,\varepsilon}[(y - \hat{f}_D(x))^2] = \sigma^2 + (\mathbb{E}_D[\hat{f}_D(x)] - f(x))^2 + \text{Var}_D(\hat{f}_D(x))$. Which term can a better learner not reduce?

4. For the one-dimensional model $y_n = wx_n + \varepsilon_n$ with fixed inputs, show difficulty ·
that the least-squares estimator $\hat{w} = \sum_n x_n y_n / s$ satisfies $\hat{w} = w + \sum_n x_n \varepsilon_n / s$, that it is unbiased, and that $\text{Var}(\hat{w}) = \sigma^2 / s$.

5. For ridge, $\hat{w}_\lambda = \sum_n x_n y_n / (s + \lambda)$, show that $\hat{w}_\lambda = \frac{s}{s + \lambda} \hat{w}$, and compute its bias, variance and MSE as functions of λ . difficulty ..

6. Find the λ^* that minimises $\text{MSE}(\hat{w}_\lambda)$ and the minimum value. Show that $\lambda^* > 0$ always, so some ridge penalty beats least squares, and evaluate for $w = 1$, $\sigma^2 = 1$, $s = 4$. difficulty ...

7. Show that $\text{bias}(\hat{w}_\lambda)^2$ is increasing in λ and $\text{Var}(\hat{w}_\lambda)$ is decreasing, by computing their derivatives, and find their limits as $\lambda \rightarrow \infty$. At what λ are the two terms equal? difficulty ..

8. B predictors $\hat{f}_1(x), \dots, \hat{f}_B(x)$ each have mean m , variance v and pairwise correlation ρ . Show that their average $\bar{f}(x) = \frac{1}{B} \sum_b \hat{f}_b(x)$ has mean m , so the bias is unchanged, and variance $\rho v + \frac{(1 - \rho)v}{B}$. What are the two limits $\rho = 0$ and $B \rightarrow \infty$? difficulty ..

9. For Gaussian data consider the estimators cS of σ^2 , where $S = \sum_n (x_n - \bar{x})^2$. Compute bias, variance and MSE as functions of c , show that the MSE is minimised at $c^* = 1/(N + 1)$, and compare the MSEs of $S/(N - 1)$ (unbiased), S/N (maximum likelihood) and $S/(N + 1)$ for $N = 5$. difficulty ...

10. Training inputs lie on a grid with spacing h , $f(x) = ax^2$, and k -nearest-neighbour regression with $k = 2m + 1$ uses the m grid points on each side of a query x that is itself a grid point. Show that the prediction has variance σ^2/k and bias $ah^2 \frac{k^2 - 1}{12}$, write the MSE above the noise floor, and find the best k among $\{1, 3, 5, 7, 9\}$ for $a = 1$, $\sigma^2 = 1$, $h = \frac{1}{4}$. difficulty ...

Answers

1. $\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$
2. $\bar{x}$: bias 0, variance σ^2/N , $\text{MSE } \sigma^2/N$; $c\bar{x}$: bias $(c-1)\mu$, variance $c^2\sigma^2/N$, $\text{MSE } (1-c)^2\mu^2 + c^2\sigma^2/N$, minimised at $c^* = \frac{\mu^2}{\mu^2 + \sigma^2/N}$; for $\mu = 1, \sigma^2 = 1, N = 4$: $c^* = 0.8$ and $\text{MSE } 0.2$ against 0.25
3. $\mathbb{E}[(y - \hat{f}_D(x))^2] = \sigma^2 + \text{bias}^2 + \text{variance}$, with $\text{bias} = \mathbb{E}_D[\hat{f}_D(x)] - f(x)$ and $\text{variance} = \text{Var}_D(\hat{f}_D(x))$; the σ^2 cannot be reduced
4. $\hat{w} = w + \sum_n x_n \varepsilon_n / s$; bias = 0; $\text{Var}(\hat{w}) = \sigma^2/s$, so $\text{MSE}(\hat{w}) = \sigma^2/s$
5. $\hat{w}_\lambda = \frac{s}{s+\lambda} \hat{w}$; bias $= -\frac{\lambda w}{s+\lambda}$; $\text{Var} = \frac{\sigma^2 s}{(s+\lambda)^2}$; $\text{MSE} = \frac{\lambda^2 w^2 + \sigma^2 s}{(s+\lambda)^2}$
6. $\lambda^* = \sigma^2/w^2$ with $\text{MSE}(\lambda^*) = \frac{\sigma^2}{s + \sigma^2/w^2} < \frac{\sigma^2}{s}$; for $w = 1, \sigma^2 = 1, s = 4$: $\lambda^* = 1$, $\text{MSE } 0.2$ against 0.25
7. $\frac{d}{d\lambda} \text{bias}^2 = \frac{2\lambda s w^2}{(s+\lambda)^3} > 0$, $\frac{d}{d\lambda} \text{Var} = -\frac{2\sigma^2 s}{(s+\lambda)^3} < 0$; limits w^2 and 0; the terms are equal at $\lambda = \sigma\sqrt{s}/w$
8. $\mathbb{E}[\bar{f}] = m$ (bias unchanged); $\text{Var}(\bar{f}) = \rho v + \frac{(1-\rho)v}{B}$; for $\rho = 0$ it is v/B , and as $B \rightarrow \infty$ it tends to ρv
9. $\text{MSE}(cS) = \sigma^4[(c(N-1)-1)^2 + 2c^2(N-1)]$, minimised at $c^* = \frac{1}{N+1}$ with $\text{MSE} = \frac{2\sigma^4}{N+1}$; for $N = 5$: $S/(N-1)$ scores $0.5\sigma^4$, S/N scores $0.36\sigma^4$, $S/(N+1)$ scores $0.333\sigma^4$
10. $\text{Var} = \sigma^2/k$, bias $= ah^2(k^2-1)/12$, $\text{MSE} - \sigma^2 = \frac{\sigma^2}{k} + a^2 h^4 \frac{(k^2-1)^2}{144}$; for $a = 1, \sigma^2 = 1, h = \frac{1}{4}$ the best of $\{1, 3, 5, 7, 9\}$ is $k = 7$, with 0.2054

Worked solutions

Problem 1

Show that for any estimator $\hat{\theta}$ of a fixed θ , $\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$.

Let $m = \mathbb{E}[\hat{\theta}]$. Then $\hat{\theta} - \theta = (\hat{\theta} - m) + (m - \theta)$.

Add and subtract the mean of the estimator; the second bracket is the bias, a constant.

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - m)^2 + 2(\hat{\theta} - m)(m - \theta) + (m - \theta)^2.$$

Expand the square.

$$\mathbb{E}[(\hat{\theta} - m)(m - \theta)] = (m - \theta) \mathbb{E}[\hat{\theta} - m] = (m - \theta) \cdot 0.$$

$m - \theta$ is a constant and comes out; $\mathbb{E}[\hat{\theta}] - m = 0$ by the definition of m .

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - m)^2] + (m - \theta)^2 = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

Take expectations of step 2; the cross term vanishes by step 3; the remaining terms are the definitions.

$$\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

The error splits into a systematic part, how far the average estimate sits from the truth, and a random part, how much the estimate moves from sample to sample, and the two add because the cross term averages out. Nothing about the distribution was used, only that θ is fixed and $\hat{\theta}$ has a mean and a variance. The decomposition is an identity, not an inequality, so lowering the MSE means lowering the sum, and Problems 2, 6 and 9 all do it by accepting some bias for a larger drop in variance; Mistake 1 refuses the trade.

Problem 2

For an i.i.d. sample with mean μ and variance σ^2 , compute the bias, variance and MSE of the sample mean $\bar{x}$. Then do the same for the shrunk estimator $c\bar{x}$ with $0 \leq c \leq 1$, find the c^* that minimises its MSE, and evaluate for $\mu = 1$, $\sigma^2 = 1$, $N = 4$.

$$\mathbb{E}[\bar{x}] = \frac{1}{N} \sum_n \mathbb{E}[x_n] = \mu, \text{ so } \text{bias}(\bar{x}) = 0.$$

Linearity of expectation; every x_n has mean μ .

$$\text{Var}(\bar{x}) = \frac{1}{N^2} \sum_n \text{Var}(x_n) = \frac{\sigma^2}{N}, \text{ so}$$

$$\text{MSE}(\bar{x}) = \frac{\sigma^2}{N}.$$

Independence makes the variance of the sum the sum of the variances;

$\text{Var}(aX) = a^2 \text{Var}(X)$ with $a = 1/N$;

Problem 1 with zero bias.

$$\mathbb{E}[c\bar{x}] = c\mu, \text{ so } \text{bias}(c\bar{x}) = (c - 1)\mu;$$

$$\text{Var}(c\bar{x}) = c^2 \sigma^2 / N.$$

Scaling by the constant c scales the mean by c and the variance by c^2 .

$$\text{MSE}(c\bar{x}) = (1 - c)^2 \mu^2 + c^2 \frac{\sigma^2}{N}.$$

Problem 1.

$$\begin{aligned} \frac{d}{dc} \text{MSE} &= -2(1 - c)\mu^2 + 2c \frac{\sigma^2}{N} = 0 \iff \\ c^* &= \frac{\mu^2}{\mu^2 + \sigma^2/N}, \text{ a minimum since the} \\ &\text{second derivative } 2\mu^2 + 2\sigma^2/N > 0. \end{aligned}$$

Differentiate the quadratic in c and solve; it is convex, so the stationary point is the minimum.

$$\mu = 1, \sigma^2 = 1, N = 4: c^* = \frac{1}{1 + 1/4} = 0.8,$$

$$\text{MSE}(0.8\bar{x}) = 0.04 \cdot 1 + 0.64 \cdot 0.25 =$$

$$0.2 < 0.25 = \text{MSE}(\bar{x}).$$

Step 5, then step 4 with $c = 0.8$: bias -0.2 , variance 0.16 .

$\bar{x}$: **bias** 0, **variance** σ^2/N , **MSE** σ^2/N ;
 $c\bar{x}$: **bias** $(c-1)\mu$, **variance** $c^2\sigma^2/N$,
MSE $(1-c)^2\mu^2 + c^2\sigma^2/N$, **minimised**
at $c^* = \frac{\mu^2}{\mu^2 + \sigma^2/N}$; **for** $\mu = 1, \sigma^2 = 1$,
 $N = 4$: $c^* = 0.8$ **and MSE** 0.2 **against**
0.25

Shrinking towards zero trades a bias of 0.2 for a variance cut from 0.25 to 0.16, and wins. The catch is that c^* depends on μ , the thing being estimated, so it is not an estimator one can use directly; it says the unbiased choice is not the floor, and that any prior knowledge of the size of μ (here, that μ^2 is comparable to σ^2/N) can be spent. Ridge regression in Problem 6 is the same calculation with the same answer, and the Bayesian posterior mean under a Gaussian prior lands on exactly $c^*\bar{x}$ with μ^2 replaced by the prior variance.

Problem 3

For a fixed test input x with $y = f(x) + \varepsilon$ and a learner $\hat{f}_D$, show that $\mathbb{E}_{D,\varepsilon}[(y - \hat{f}_D(x))^2] = \sigma^2 + (\mathbb{E}_D[\hat{f}_D(x)] - f(x))^2 + \text{Var}_D(\hat{f}_D(x))$. Which term can a better learner not reduce?

$$y - \hat{f}_D(x) = \varepsilon + (f(x) - \hat{f}_D(x)).$$

Substitute $y = f(x) + \varepsilon$ and regroup.

$$(y - \hat{f}_D(x))^2 = \varepsilon^2 + 2\varepsilon(f(x) - \hat{f}_D(x)) + (f(x) - \hat{f}_D(x))^2.$$

Expand the square.

$$\mathbb{E}[\varepsilon^2] = \sigma^2 \text{ and } \mathbb{E}[\varepsilon(f(x) - \hat{f}_D(x))] = \mathbb{E}[\varepsilon] \mathbb{E}_D[f(x) - \hat{f}_D(x)] = 0.$$

ε has mean zero and variance σ^2 ; the test noise is independent of the training set D , so the expectation of the product is the product of the expectations (the variance page).

$$\mathbb{E}_D[(f(x) - \hat{f}_D(x))^2] = (\mathbb{E}_D[\hat{f}_D(x)] - f(x))^2 + \text{Var}_D(\hat{f}_D(x)).$$

Problem 1 with $\hat{\theta} = \hat{f}_D(x)$, an estimator of the fixed number $\theta = f(x)$.

$\mathbb{E}[(y - \hat{f}_D(x))^2] = \sigma^2 + \text{bias}^2 + \text{variance}$,
with $\text{bias} = \mathbb{E}_D[\hat{f}_D(x)] - f(x)$ **and**
 $\text{variance} = \text{Var}_D(\hat{f}_D(x))$; **the σ^2 cannot**
be reduced

The expected test error at a point has a floor, the noise variance, that no learner reaches: even $\hat{f}_D = f$ exactly scores σ^2 (Mistake 4). Above the floor sit the two terms of Problem 1 for the learner regarded as an estimator of $f(x)$; averaging over x gives the usual statement about test error. Step 3 used that the test noise is independent of D , and that is what fails on the training set, where the same ε_n sits in y_n and in $\hat{f}_D$ (Mistake 2). The check evaluates the left side exactly for the linear model of Problems 4 and 5, where $\hat{f}_D(x) = \hat{w} x$ is Gaussian, and compares it with the three terms.

Problem 4

For the one-dimensional model $y_n = wx_n + \varepsilon_n$ with fixed inputs, show that the least-squares estimator $\hat{w} = \sum_n x_n y_n / s$ satisfies $\hat{w} = w + \sum_n x_n \varepsilon_n / s$, that it is unbiased, and that $\text{Var}(\hat{w}) = \sigma^2 / s$.

$$\sum_n x_n y_n = \sum_n x_n (wx_n + \varepsilon_n) = ws + \sum_n x_n \varepsilon_n.$$

Substitute the model and use $\sum_n x_n^2 = s$.

$$\hat{w} = \frac{ws + \sum_n x_n \varepsilon_n}{s} = w + \frac{\sum_n x_n \varepsilon_n}{s}.$$

Divide by s .

$$\mathbb{E}[\hat{w}] = w + \frac{\sum_n x_n \mathbb{E}[\varepsilon_n]}{s} = w.$$

The x_n are fixed and every ε_n has mean zero.

$$\text{Var}(\hat{w}) = \frac{1}{s^2} \sum_n x_n^2 \text{Var}(\varepsilon_n) = \frac{\sigma^2 s}{s^2} = \frac{\sigma^2}{s}.$$

Independent noise: the variance of $\sum_n x_n \varepsilon_n$ is $\sum_n x_n^2 \sigma^2$; then $\text{Var}(aX) = a^2 \text{Var}(X)$ with $a = 1/s$.

$$\hat{w} = w + \sum_n x_n \varepsilon_n / s; \text{bias} = 0; \text{Var}(\hat{w}) = \sigma^2 / s, \text{ so } \text{MSE}(\hat{w}) = \sigma^2 / s$$

The estimator is the truth plus a weighted average of the noise, and its variance falls as $s = \sum_n x_n^2$ grows: more points, or points farther from the origin, pin the slope down. This is the scalar case of the least-squares page's $\text{Cov}(\hat{w}) = \sigma^2 (X^\top X)^{-1}$, with $X^\top X = s$. The regression page shows that least squares is also the maximum-likelihood estimator under Gaussian noise, and the MLE page's variance estimator is biased for the same structural reason that Problem 9 revisits; the slope estimator itself is not.

Problem 5

For ridge, $\hat{w}_\lambda = \sum_n x_n y_n / (s + \lambda)$, show that $\hat{w}_\lambda = \frac{s}{s + \lambda} \hat{w}$, and compute its bias, variance and MSE as functions of λ .

$$\hat{w}_\lambda = \frac{\sum_n x_n y_n}{s + \lambda} = \frac{s}{s + \lambda} \cdot \frac{\sum_n x_n y_n}{s} = \frac{s}{s + \lambda} \hat{w}.$$

Multiply and divide by s .

$$\mathbb{E}[\hat{w}_\lambda] = \frac{s}{s + \lambda} w, \text{ so } \text{bias}(\hat{w}_\lambda) = \frac{s}{s + \lambda} w - w = -\frac{\lambda}{s + \lambda} w.$$

Problem 4's $\mathbb{E}[\hat{w}] = w$ scaled by the constant; $\frac{s}{s + \lambda} - 1 = -\frac{\lambda}{s + \lambda}$.

$$\text{Var}(\hat{w}_\lambda) = \left(\frac{s}{s + \lambda} \right)^2 \frac{\sigma^2}{s} = \frac{\sigma^2 s}{(s + \lambda)^2}.$$

$\text{Var}(aX) = a^2 \text{Var}(X)$ with Problem 4's variance.

$$\text{MSE}(\hat{w}_\lambda) = \frac{\lambda^2 w^2}{(s + \lambda)^2} + \frac{\sigma^2 s}{(s + \lambda)^2} = \frac{\lambda^2 w^2 + \sigma^2 s}{(s + \lambda)^2}.$$

Problem 1; common denominator.

$$\hat{w}_\lambda = \frac{s}{s+\lambda} \hat{w}; \text{bias} = -\frac{\lambda w}{s+\lambda}; \text{Var} = \frac{\sigma^2 s}{(s+\lambda)^2}; \text{MSE} = \frac{\lambda^2 w^2 + \sigma^2 s}{(s+\lambda)^2}$$

Ridge is Problem 2's shrinkage with $c = s/(s+\lambda)$: the estimate is pulled towards zero by a factor that the data size s and the penalty λ fight over, the bias is the part of w lost to the pull, and the variance is cut by c^2 . At $\lambda = 0$ everything reduces to Problem 4; as $\lambda \rightarrow \infty$ the estimate goes to 0, with bias $-w$ and no variance. The variance is $\sigma^2 s/(s+\lambda)^2$ and not $\sigma^2/(s+\lambda)$ (Mistake 3), a distinction that matters for everything in Problems 6 and 7.

Problem 6

Find the λ^* that minimises $\text{MSE}(\hat{w}_\lambda)$ and the minimum value. Show that $\lambda^* > 0$ always, so some ridge penalty beats least squares, and evaluate for $w = 1, \sigma^2 = 1, s = 4$.

$$\text{MSE}(\lambda) = \frac{u(\lambda)}{v(\lambda)} \text{ with } u = \lambda^2 w^2 + \sigma^2 s \text{ and } v = (s + \lambda)^2; u' = 2\lambda w^2 \text{ and } v' = 2(s + \lambda).$$

Name the pieces for the quotient rule.

$$\frac{d}{d\lambda} \text{MSE} = \frac{u'v - uv'}{v^2} = \frac{2(s + \lambda)[\lambda w^2(s + \lambda) - \lambda^2 w^2 - \sigma^2 s]}{(s + \lambda)^4} = \frac{2s(\lambda w^2 - \sigma^2)}{(s + \lambda)^3}.$$

Quotient rule; factor $2(s + \lambda)$ from the numerator; inside the bracket $\lambda w^2 s + \lambda^2 w^2 - \lambda^2 w^2 - \sigma^2 s = s(\lambda w^2 - \sigma^2)$.

The derivative is negative for $\lambda < \sigma^2/w^2$ and positive for $\lambda > \sigma^2/w^2$, so $\lambda^* = \frac{\sigma^2}{w^2}$ is the minimum.

The sign of step 2 is the sign of $\lambda w^2 - \sigma^2$; the function decreases then increases.

$$\text{MSE}(\lambda^*) = \frac{\sigma^4/w^2 + \sigma^2 s}{(s + \sigma^2/w^2)^2} = \frac{\sigma^2(s + \sigma^2/w^2)}{(s + \sigma^2/w^2)^2} = \frac{\sigma^2}{s + \sigma^2/w^2}.$$

Substitute λ^* ; the numerator is σ^2 times the base of the denominator.

$$\lambda^* = \sigma^2/w^2 > 0 \text{ whenever } \sigma^2 > 0, \text{ and}$$

$$\frac{\sigma^2}{s + \sigma^2/w^2} < \frac{\sigma^2}{s} = \text{MSE}(\hat{w}).$$

A positive number divided by w^2 ; a larger denominator.

$w = 1, \sigma^2 = 1, s = 4$: $\lambda^* = 1$, $\text{MSE}(\lambda^*) = \frac{1}{4+1} = 0.2$ against $\text{MSE}(\hat{w}) = 0.25$; at $\lambda^* = 1$ the bias is -0.2 and the variance $4/25 = 0.16$.

Steps 3 and 4; Problem 5's bias and variance at $\lambda = 1$.

$\lambda^* = \sigma^2/w^2$ **with** $\text{MSE}(\lambda^*) = \frac{\sigma^2}{s + \sigma^2/w^2} < \frac{\sigma^2}{s}$; **for** $w = 1, \sigma^2 = 1, s = 4$: $\lambda^* = 1$, **MSE 0.2 against 0.25**

The best penalty is the noise-to-signal ratio σ^2/w^2 , and it does not depend on s : more data does not change the right λ , it makes the shrinkage factor $s/(s + \lambda)$ closer to 1 so that the penalty matters less. The numbers are Problem 2's exactly, because $\hat{w}_\lambda = c\hat{w}$ with $c = s/(s + \lambda)$ and λ^* corresponds to $c^* = w^2/(w^2 + \sigma^2/s)$. In the Bayesian reading, $\lambda = \sigma^2/\tau^2$ is the ridge penalty that makes the posterior mean under the prior $w \sim \mathcal{N}(0, \tau^2)$, and λ^* is the prior whose variance matches w^2 . As in Problem 2, λ^* needs w , so in practice it is chosen by cross-validation, which estimates the MSE curve of Problem 7 from the data.

Problem 7

Show that $\text{bias}(\hat{w}_\lambda)^2$ is increasing in λ and $\text{Var}(\hat{w}_\lambda)$ is decreasing, by computing their derivatives, and find their limits as $\lambda \rightarrow \infty$. At what λ are the two terms equal?

$$\text{bias}^2 = \frac{\lambda^2 w^2}{(s + \lambda)^2} \text{ and } \frac{d}{d\lambda} \text{bias}^2 = w^2 \cdot \frac{2\lambda(s + \lambda)^2 - \lambda^2 \cdot 2(s + \lambda)}{(s + \lambda)^4} = \frac{2\lambda s w^2}{(s + \lambda)^3} > 0$$

for $\lambda > 0$.

Quotient rule; factor $2\lambda(s + \lambda)$ from the numerator, leaving $(s + \lambda) - \lambda = s$.

$$\text{Var} = \frac{\sigma^2 s}{(s + \lambda)^2} \text{ and } \frac{d}{d\lambda} \text{Var} = -\frac{2\sigma^2 s}{(s + \lambda)^3} < 0.$$

Power rule on $(s + \lambda)^{-2}$.

As $\lambda \rightarrow \infty$: $\text{bias}^2 \rightarrow w^2$ and $\text{Var} \rightarrow 0$.

$\lambda^2/(s + \lambda)^2 \rightarrow 1$ and $s/(s + \lambda)^2 \rightarrow 0$.

$$\text{bias}^2 = \text{Var} \iff \lambda^2 w^2 = \sigma^2 s \iff \lambda = \frac{\sigma \sqrt{s}}{w}.$$

Equate the numerators, which share the denominator; take the positive root.

$$\frac{d}{d\lambda} \text{bias}^2 = \frac{2\lambda s w^2}{(s + \lambda)^3} > 0, \frac{d}{d\lambda} \text{Var} = -\frac{2\sigma^2 s}{(s + \lambda)^3} < 0; \text{limits } w^2 \text{ and } 0; \text{ the terms are equal at } \lambda = \sigma \sqrt{s}/w$$

This is the picture behind every regularisation plot: a rising bias curve, a falling variance curve and a U-shaped sum. Adding the two derivatives gives $\frac{2s(\lambda w^2 - \sigma^2)}{(s + \lambda)^3}$, Problem 6's derivative, and the sum is minimised where the two slopes cancel, at $\lambda^* = \sigma^2/w^2$, which is not where the curves cross: for Problem 6's numbers they cross at $\lambda = 2$ while the minimum is at $\lambda = 1$, where the variance term (0.16) is still four times the squared bias (0.04). The same shape appears with model size, training time and the number of neighbours (Problem 10) as the knob.

Problem 8

B predictors $\hat{f}_1(x), \dots, \hat{f}_B(x)$ each have mean m , variance v and pairwise correlation ρ . Show that their average $\bar{f}(x) = \frac{1}{B} \sum_b \hat{f}_b(x)$ has mean m , so the bias is unchanged, and variance $\rho v + \frac{(1 - \rho)v}{B}$. What are the two limits $\rho = 0$ and $B \rightarrow \infty$?

$$\mathbb{E}[\bar{f}] = \frac{1}{B} \sum_b \mathbb{E}[\hat{f}_b] = \frac{1}{B} \cdot Bm = m.$$

Linearity; every predictor has the same mean. The bias $m - f(x)$ is therefore the same as each member's.

Let $z = (\hat{f}_1, \dots, \hat{f}_B)^\top$ with covariance Σ , $\Sigma_{bb} = v$ and $\Sigma_{bb'} = \rho v$ for $b \neq b'$. Then $\bar{f} = \frac{1}{B} \mathbf{1}^\top z$ and $\text{Var}(\bar{f}) = \frac{1}{B^2} \mathbf{1}^\top \Sigma \mathbf{1}$.

The variance page's $\text{Cov}(Az) = A \Sigma A^\top$ with the $1 \times B$ matrix $A = \frac{1}{B} \mathbf{1}^\top$.

$$\mathbf{1}^\top \Sigma \mathbf{1} = \sum_{b,b'} \Sigma_{bb'} = Bv + B(B-1)\rho v.$$

B diagonal entries equal to v and $B(B-1)$ off-diagonal entries equal to ρv .

$$\text{Var}(\bar{f}) = \frac{Bv + B(B-1)\rho v}{B^2} = \frac{v + (B-1)\rho v}{B} = \rho v + \frac{(1-\rho)v}{B}.$$

Divide by B^2 ; split $(B-1)\rho v = B\rho v - \rho v$.

$\mathbb{E}[\bar{f}] = m$ (**bias unchanged**); $\text{Var}(\bar{f}) = \rho v + \frac{(1-\rho)v}{B}$; **for $\rho = 0$ it is v/B , and as $B \rightarrow \infty$ it tends to ρv**

Averaging models attacks only the variance term of Problem 3, and only the part of it that is not shared: with independent members the variance falls like $1/B$ (the variance page's sample mean again), and with correlated members it stops at the floor ρv however many are averaged (Mistake 5). Bagging trains members on bootstrap resamples to make them differ, and random forests decorrelate them further by restricting each split to a random subset of features, lowering ρ at the cost of a little bias; the Jensen page's ambiguity decomposition is the same conclusion reached for one dataset rather than in expectation. Boosting is the opposite strategy: it attacks the bias term by fitting each new model to the residual of the current average.

Problem 9

For Gaussian data consider the estimators cS of σ^2 , where $S = \sum_n (x_n - \bar{x})^2$. Compute bias, variance and MSE as functions of c , show that the MSE is minimised at $c^* = 1/(N+1)$, and

compare the MSEs of $S/(N - 1)$ (unbiased), S/N (maximum likelihood) and $S/(N + 1)$ for $N = 5$.

$$\mathbb{E}[S] = (N - 1)\sigma^2 \text{ and } \text{Var}(S) = 2(N - 1)\sigma^4.$$

S/σ^2 is chi-square with $N - 1$ degrees of freedom, with mean $N - 1$ and variance $2(N - 1)$; scaling by σ^2 scales the mean by σ^2 and the variance by σ^4 .

$$\text{bias}(cS) = c(N - 1)\sigma^2 - \sigma^2 = (c(N - 1) - 1)\sigma^2 \text{ and } \text{Var}(cS) = 2c^2(N - 1)\sigma^4.$$

Scale step 1 by c ; subtract the target σ^2 .

$$\text{MSE}(cS) = \sigma^4 \left[(c(N - 1) - 1)^2 + 2c^2(N - 1) \right].$$

Problem 1.

$$\frac{d}{dc} \text{MSE} = \sigma^4 \left[2(N - 1)(c(N - 1) - 1) + 4c(N - 1) \right] = 2(N - 1)\sigma^4 [c(N - 1) - 1 + 2c] = 2(N - 1)\sigma^4 [c(N + 1) - 1].$$

Differentiate the quadratic in c ; collect the terms in c .

$$c^* = \frac{1}{N + 1}, \text{ a minimum since the MSE is a convex quadratic in } c; \text{ there bias} = -\frac{2\sigma^2}{N + 1}, \text{ Var} = \frac{2(N - 1)\sigma^4}{(N + 1)^2} \text{ and MSE} = \frac{\sigma^4(4 + 2(N - 1))}{(N + 1)^2} = \frac{2\sigma^4}{N + 1}.$$

Set step 4 to zero; substitute c^* into step 2 and add.

$$c = \frac{1}{N - 1}: \text{bias } 0, \text{ MSE} = \frac{2\sigma^4}{N - 1}. \quad c = \frac{1}{N}: \text{bias } -\frac{\sigma^2}{N}, \text{ MSE} = \sigma^4 \left[\frac{1}{N^2} + \frac{2(N - 1)}{N^2} \right] = \frac{(2N - 1)\sigma^4}{N^2}.$$

Step 3 at the two values; $(\frac{N-1}{N} - 1)^2 = \frac{1}{N^2}$.

$N = 5$, in units of σ^4 : $S/4$ gives $\frac{2}{4} = 0.5$; $S/5$ gives $\frac{9}{25} = 0.36$; $S/6$ gives $\frac{2}{6} \approx 0.333$.

Steps 5 and 6 with $N = 5$.

$\text{MSE}(cS) = \sigma^4 [(c(N-1) - 1)^2 + 2c^2(N-1)]$, **minimised at** $c^* = \frac{1}{N+1}$
with $\text{MSE} = \frac{2\sigma^4}{N+1}$; **for** $N = 5$: $S/(N-1)$ **scores** $0.5\sigma^4$, S/N **scores** $0.36\sigma^4$,
 $S/(N+1)$ **scores** $0.333\sigma^4$

The unbiased estimator has the largest error of the three: its lack of bias is bought with variance, and the maximum-likelihood $1/N$ already does better; the best divisor is $N+1$, which nobody uses because the gain over $1/N$ is small and the argument requires Gaussian data. The MLE page's Problem 5 explains the bias of S/N ; this problem shows that bias was not the thing to minimise. The check evaluates $\mathbb{E}[S]$ and $\mathbb{E}[S^2]$ exactly by quadrature for small N , so the chi-square facts are verified rather than assumed.

Problem 10

Training inputs lie on a grid with spacing h , $f(x) = ax^2$, and k -nearest-neighbour regression with $k = 2m + 1$ uses the m grid points on each side of a query x that is itself a grid point.

Show that the prediction has variance σ^2/k and bias $ah^2 \frac{k^2 - 1}{12}$, write the MSE above the noise floor, and find the best k among $\{1, 3, 5, 7, 9\}$ for $a = 1$, $\sigma^2 = 1$, $h = \frac{1}{4}$.

$$\hat{f}(x) = \frac{1}{k} \sum_{j=-m}^m (f(x + jh) + \varepsilon_j), \text{ so}$$

$$\text{Var}(\hat{f}(x)) = \frac{1}{k^2} \cdot k\sigma^2 = \frac{\sigma^2}{k}.$$

The k neighbours are the grid points $x + jh$; their noises are independent with variance σ^2 , and the average of k of them has variance σ^2/k .

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{k} \sum_{j=-m}^m a(x + jh)^2 = ax^2 + \frac{2axh}{k} \sum_j j + \frac{ah^2}{k} \sum_j j^2 = ax^2 + \frac{ah^2}{k} \sum_{j=-m}^m j^2.$$

Expand the square; $\sum_{j=-m}^m j = 0$ by symmetry.

$$\sum_{j=-m}^m j^2 = 2 \cdot \frac{m(m+1)(2m+1)}{6} = \frac{m(m+1)k}{3}, \text{ so bias} = \mathbb{E}[\hat{f}(x)] - ax^2 = \frac{ah^2 m(m+1)}{3}.$$

The sum of the first m squares, doubled for the negative side; $2m+1 = k$ cancels the $1/k$.

$$m = \frac{k-1}{2} \text{ gives } m(m+1) = \frac{(k-1)(k+1)}{4} = \frac{k^2-1}{4}, \text{ so bias} = ah^2 \frac{k^2-1}{12}.$$

Substitute and simplify.

$$\text{MSE} - \sigma^2 = \frac{\sigma^2}{k} + a^2 h^4 \frac{(k^2-1)^2}{144}.$$

Problem 3's bias squared plus variance.

With $a = 1$, $\sigma^2 = 1$, $h = \frac{1}{4}$ ($h^4 = \frac{1}{256}$): $k = 1$: 1; $k = 3$: $0.3333 + \frac{64}{36864} = 0.3351$; $k = 5$: $0.2 + \frac{576}{36864} = 0.2156$; $k = 7$: $0.1429 + \frac{2304}{36864} = 0.2054$; $k = 9$: $0.1111 + \frac{6400}{36864} = 0.2847$.

Step 5 term by term; $144 \cdot 256 = 36864$.

$\text{Var} = \sigma^2/k$, $\text{bias} = ah^2(k^2-1)/12$,
 $\text{MSE} - \sigma^2 = \frac{\sigma^2}{k} + a^2 h^4 \frac{(k^2-1)^2}{144}$; **for** $a = 1$, $\sigma^2 = 1$, $h = \frac{1}{4}$ **the best of** $\{1, 3, 5, 7, 9\}$ **is** $k = 7$, **with** 0.2054

$k = 1$ has no bias and all the variance; growing k averages the noise down as $1/k$ but reaches farther from x , where the curvature a of f makes the neighbours' average drift above $f(x)$, and that bias grows like k^2 . The best k depends on everything: more curvature or coarser data (larger ah^2) wants a smaller k , and more noise wants a larger one. With $h = 1$ the same numbers give $k = 3$, and with $h = \frac{1}{16}$ they give $k = 19$: denser data allows more smoothing, which is the general rule that the optimal amount of regularisation falls, but does not vanish, as the dataset grows. The bias here is pure curvature, which is why k -NN is unbiased for linear f on a symmetric neighbourhood and why the symmetry assumption fails at the edge of the data.

Where this goes wrong

1. Preferring the unbiased estimator by reflex

Unbiased sounds like correct, and an estimator that is biased on purpose sounds like a trick.

$\text{MSE}(\hat{\theta}) = \text{bias}^2 + \text{Var}$, and $\bar{x}$ has bias 0

Right so far: Problems 1 and 2.

“Zero bias means the error is as small as it can be.”

The habit that causes the mistake: zero bias sets one term of the sum to zero and says nothing about the other.

$\text{MSE}(c\bar{x}) \geq \text{MSE}(\bar{x})$ for every $c \neq 1$ ✗

For $\mu = 1$, $\sigma^2 = 1$, $N = 4$, $c = 0.8$ gives $0.2 < 0.25$ (Problem 2); ridge at $\lambda^* = \sigma^2/w^2$ beats least squares for every w and s (Problem 6); and $S/(N+1)$ beats the unbiased $S/(N-1)$ for every N (Problem 9). What is true is that among unbiased estimators the one with the smallest variance is best, and that is a different and smaller competition. The reflex has a real basis, though: shrinkage needs to know which way to shrink and how far, and when nothing is known about θ the unbiased estimator is the one that cannot be fooled. Regularisation is exactly the act of putting that knowledge, usually “the true parameters are not huge”, into the estimator.

2. Applying the decomposition to the training error

The three-term formula is about expected squared error, and the error that is to hand is the one on the training set.

$\mathbb{E}[(y - \hat{f}_D(x))^2] = \sigma^2 + \text{bias}^2 + \text{Var}$ for a test point (x, y)

Right so far: Problem 3.

“The training points are points too, so the same formula gives the expected training error.”

The habit that causes the mistake: Problem 3's step 3 dropped the cross term because the test noise ε is independent of D ; a training point's noise ε_n is part of D .

$$\mathbb{E}[(y_n - \hat{f}_D(x_n))^2] = \sigma^2 + \text{bias}(x_n)^2 + \text{Var}(x_n) \text{ for a training point } x_n \quad \times$$

The cross term is $-2 \mathbb{E}[\varepsilon_n(\hat{f}_D(x_n) - f(x_n))]$, and for Problem 4's least squares, $\hat{f}_D(x_n) - f(x_n) = x_n \sum_m x_m \varepsilon_m / s$, so it equals $-2x_n^2 \sigma^2 / s < 0$: the fit leans towards its own noise. Averaged over the N training points the expected training error is $\sigma^2 - \sigma^2/N$, below the noise floor that no test error can beat, and the check confirms $\sigma^2(N-1)/N$ exactly. The general version is $\sigma^2(N-p)/N$ for a p -parameter linear model, the reason the MLE page's $\hat{\sigma}^2$ is biased low, and the reason training error underestimates test error by an amount that grows with the number of parameters.

3. Guessing the ridge variance by analogy

Least squares has variance σ^2/s , ridge replaces s by $s + \lambda$ everywhere else, and the variance is written the same way.

$$\hat{w} = \sum_n x_n y_n / s \text{ has } \text{Var}(\hat{w}) = \sigma^2 / s$$

Right so far: Problem 4.

“Ridge just replaces s by $s + \lambda$.”

The shortcut that causes the mistake: it does in the estimator, and the variance is not linear in the estimator's denominator.

$$\text{Var}(\hat{w}_\lambda) = \frac{\sigma^2}{s + \lambda} \mathbf{X}$$

$\hat{w}_\lambda = \sum_n x_n y_n / (s + \lambda)$ is $\sum_n x_n \varepsilon_n / (s + \lambda)$ plus a constant, with variance $\sigma^2 s / (s + \lambda)^2$ (Problem 5): the numerator keeps the s because the noise enters through $\sum_n x_n \varepsilon_n$, whose variance is $\sigma^2 s$ regardless of λ . The two agree only at $\lambda = 0$; at Problem 6's $\lambda^* = 1$ the wrong formula gives 0.2 where the variance is 0.16, and it reports the total MSE at λ^* as 0.24 rather than 0.2, nearly erasing the gain over least squares. The matrix version has the same shape: $\text{Cov}(\hat{w}_\lambda) = \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}$, not $\sigma^2 (X^\top X + \lambda I)^{-1}$, the latter being the Bayesian posterior covariance, a different object that answers a different question.

4. Dropping the noise term

Bias and variance are the two things a model can change, and the decomposition is remembered as those two.

$$\mathbb{E}_D [(f(x) - \hat{f}_D(x))^2] = \text{bias}^2 + \text{Var}$$

Right so far: Problem 3, step 4, for the error against the noiseless $f(x)$.

“Test error is bias squared plus variance.”

The shortcut that causes the mistake: that is the error against $f(x)$; the test error is measured against $y = f(x) + \varepsilon$, and the noise does not go away because it is not the model's fault.

A learner with zero bias and zero variance has expected test error 0 ✗

It has expected test error σ^2 , the floor in Problem 3, and so does $\hat{f}_D = f$ itself. The practical damage is a target that cannot be hit: a model whose test error has stopped at σ^2 is being pushed to fit noise, and the bias–variance accounting then charges every extra bit of fit to variance. A related reading error is to take Problem 3 as saying the test error is at least σ^2 at every point; it is at least σ^2 in expectation, and Mistake 2 shows the training error sitting below it because its noise is not independent of the fit.

5. Dividing the variance of correlated models by their number

Averaging B independent measurements divides the variance by B , and the B models in an ensemble are counted as if they were independent.

$$\text{Var}(\bar{f}) = \frac{1}{B^2} \mathbf{1}^\top \Sigma \mathbf{1}$$

Right so far: Problem 8, step 2.

“Each model is its own fit, so the models are independent and $\Sigma = vI$.”

The habit that causes the mistake: models trained on resamples of the same data, with the same features and the same algorithm, make many of the same errors, and Σ has off-diagonal entries ρv with ρ typically well above zero.

$\text{Var}(\bar{f}) = \frac{v}{B}$, so a large enough ensemble has negligible variance ✗

It is $\rho v + (1 - \rho)v/B$ (Problem 8), which never falls below ρv : with $\rho = 0.5$, averaging 100 models leaves $0.505v$, almost the same as averaging 10 ($0.55v$), and the $1/B$ line predicts $0.01v$. The error compounds if the leftover variance is then blamed on bias and the members made more flexible, which raises v and often ρ too. The lever that works is ρ : random forests subsample features at every split for exactly this reason, and the variance page's sample mean under common correlation is the same formula with the same floor.

eulerpad.app/practice/bias-variance-decomposition